1) Сбор данных, реализация скриптов для чистки данных

Реализация web-crawler для сбора коллекции документов с указанного ресурса. Выбор формата хранения коллекции.

Объем собранной коллекции не менее 100 тысяч уникальных документов. Также в данной задаче необходимо реализовать механизм очистки документов от не релевантной информации, например:

- html теги
- ссылки на сторонние или внутренние ресурсы сайта
- ссылки на изображения и видео
- знаки препинания

Также данный механизм должен уметь выделять такие признаки документа как автор, тематические теги, рейтинг, репосты, дата публикации документа и другую мета информацию предоставляемую ресурсом.

После проведенной обработки размер текста для каждого документа должен быть не менее 2000 символов.

Результатом сдачи является код web-crawler, инструмент очистки текстов, исходная коллекция документов и коллекция документов после очистки.

2) Разведочный анализ данных, реализация скриптов для подготовки выборок

Проведение разведочного анализа включает в себя такие подзадачи как:

- частотный анализ коллекции(например распределение количества слов в документах)
- выбор пространства признаков с обоснование данных признаков
- визуализация данных
- подсчет статистических параметров для коллекции с использованием выбранных признаков(например среднее, стандартное отклонение, квартили и тд)
- поиск выбросов и аномалий, их анализ, выбор методов детектирования аномалий и их обработки.
- корреляционный анализ(например, построенные парные диаграммы рассеяния различных признаков, их гистограммы распределения, корреляционные матрицы, проверка гипотезы о гомоскедастичности данных и подобных ей)
- выбор целевых переменных для задач классификации документов по темам на основе тегов и предсказания рейтинга документа.
- построение распределения и анализ целевых переменных

Для полноты анализа могут быть описаны различные преобразования данных и их анализ, применение PCA, анализ feature importance, поиск ложных корреляций.

После проведения каждого анализа следует фиксировать выводы о полученных результатах.

Для сдачи необходимо представить ноутбуки в которых проведен соответствующий анализ и описаны выводы и рекомендации по работе с данной коллекцией, а также скрипты на выбранном языке программирования.

3) Классификация / кластеризация / topic modeling

В данной задаче необходимо провести три эксперимента связанных с задачами классификации текстов по темам на основе тегов, их кластеризации и тематическому моделированию. Необходимо выбрать представление данных и набор признаков(обосновать представление и выбранные признаки на основе разведочного анализа) и провести каждый эксперимент не менее чем на трех различных моделях(например для классификации наивный байес, дерево решений и метод опорных векторов, прописать обоснование выбора именного таких моделей).
Так же необходимо выбрать метрики оценки решения (для каждой из задач - классификации, кластеризации, topic modeling) и обосновать почему вы считаете что такого набора будет достаточно и он покажет качество, стабильность модели, а также позволит продемонстрировать преимущества и недостатки используемых алгоритмом.
Эксперимент должен состоять из построения датасета, разделения его на две части: тестовую и тренировочную, проведения обучения моделей и их оценки. После этого необходимо описать полученные результаты, выводы и рекомендации к ним.
Реализовать метод подбора оптимального набора гиперпараметров для моделей и аргументировать выбор конкретного алгоритма подбора оптимального набора гиперпараметров в сравнении с аналогами.
Сдавать надо будет набор ноутбуков в которых находиться код экспериментов, обоснования и выводы, а также набор скриптов с моделями и оболочку в виде сервиса с rest-api и документацией

4) Регрессия / построение рекомендации с указанием метрики

Построение программного комплекса на основе модели машинного обучения, который предоставляет возможность оценки документа и дает рекомендации по улучшению данного текста(например увеличение объема текста или добавления вводных конструкций в текст).
Например, для интерпретируемых признаков можно искать оптимальные диапазоны значений и выдавать рекомендации по улучшению целевого показателя.
Для не интерпретируемых признаков можно посмотреть на отличия текста от похожих, но с большей оценкой. И на основе таких отличий написать алгоритм формирования рекомендации.

Для успешной сдачи задачи необходимо провести следующие исследования:

- выбор пространства признаков
- построение модели регрессии по рейтингу
- интерпретация данной модели и анализ предсказания модели

Все исследования проводятся в ноутбуках и в них же фиксируются выводы по результатам исследований.
По завершению исследований необходимо реализовать программный комплекс позволяющий проводить оценку и рекомендации по улучшению текста. Желательный формат реализации в виде простого веб сервиса с user friendly интерфейсом.